AMD
Smarter Choice

2007 Hot Chips 19
AMD's Radeon™ HD 2900
2nd Generation Unified Shader Architecture

Mike Mantor
Fellow
AMD Graphics Products Group
michael.mantor@amd.com

AMD Radeon HD™ 2900 Highlights

Technology leadership

- Clock speeds – 742 MHz
- Transistor – 700 million
- Technology Process - TSMC 80nm HS
- Power ~215 W, Pin Count - 2140
- Die Size 420mm (20mm x 21mm)

Cutting-edge image quality features

- Advanced anti-aliasing and texture filtering capabilities
- Fast High Dynamic Range rendering
- Programmable Tessellation Unit

2nd generation unified architecture

- Scalar ALU design with 320 stream processing units
- 475 GigaFLOPS of (MulAdd) compute
- 47.5 GigaPixels/Sec & 742 Mtri/sec
- 106 GB/sec Bandwidth
- Optimized for Dynamic Game Computing and Accelerated Stream Processing

ATI Avivo™ HD technology

- Delivering The Ultimate Visual Experience™ For HD video
- HD display and audio connectivity
- HD DVD and Blu-Ray capable

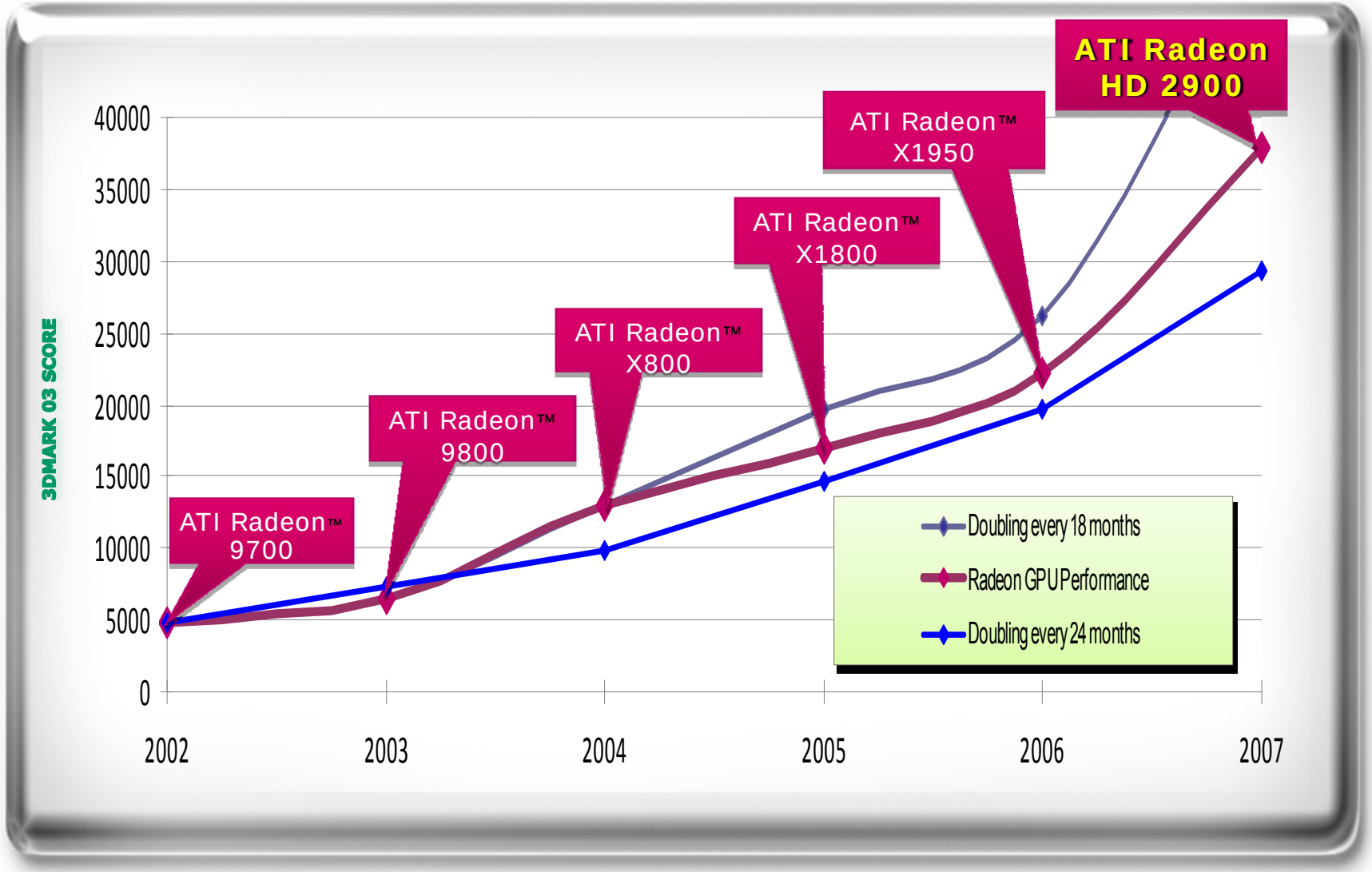
DirectX® 10

- Massive shader and geometry processing performance
- Shader Model 4.0 with Integer support
- Enabling the next generation of visual effects

Native CrossFire™ technology

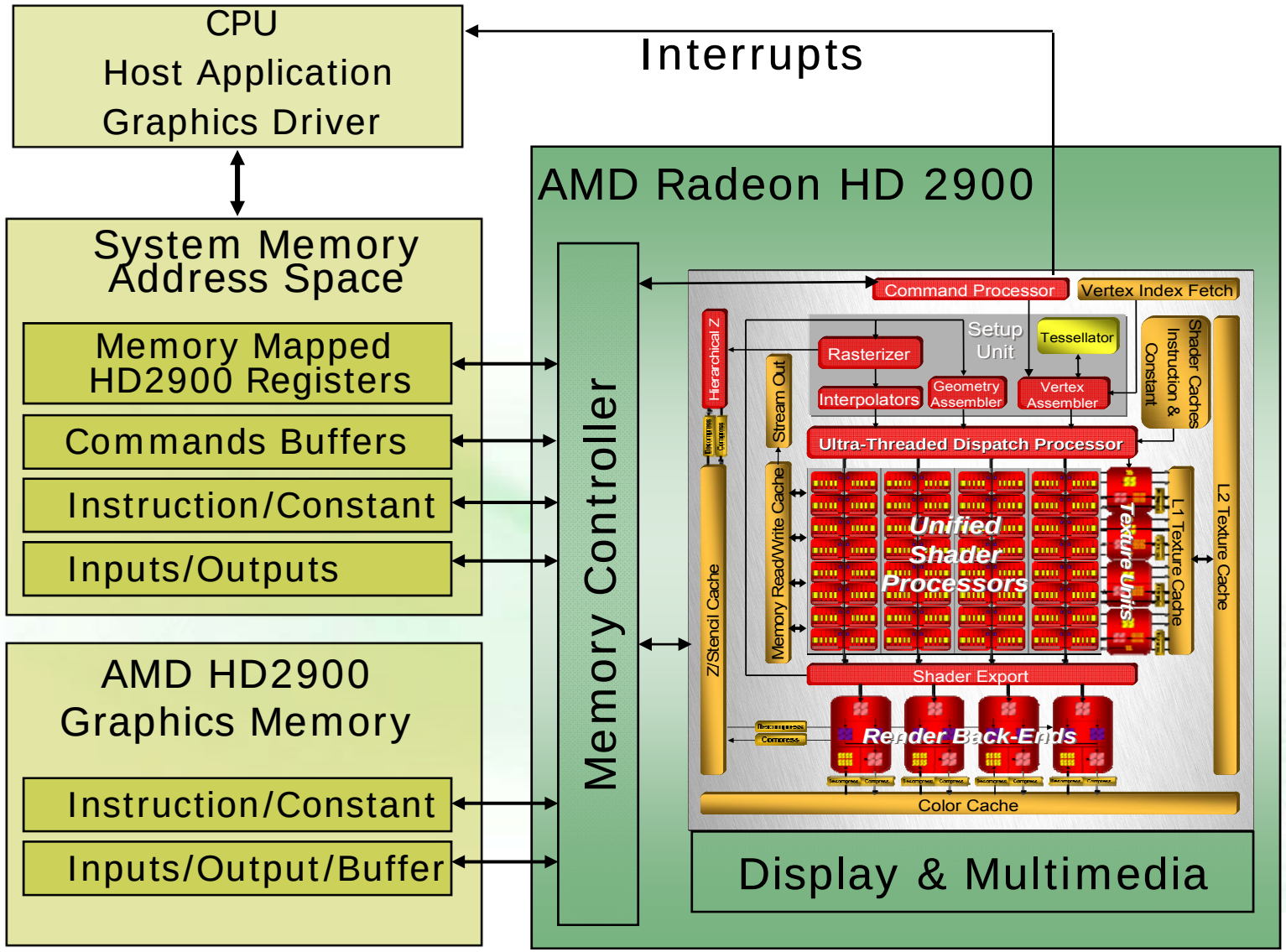
- Superior multi-GPU support
- Scales up rendering performance and image quality with 2 or more GPUs

Performance Improvements (5 years)



– Approximate Availability Dates

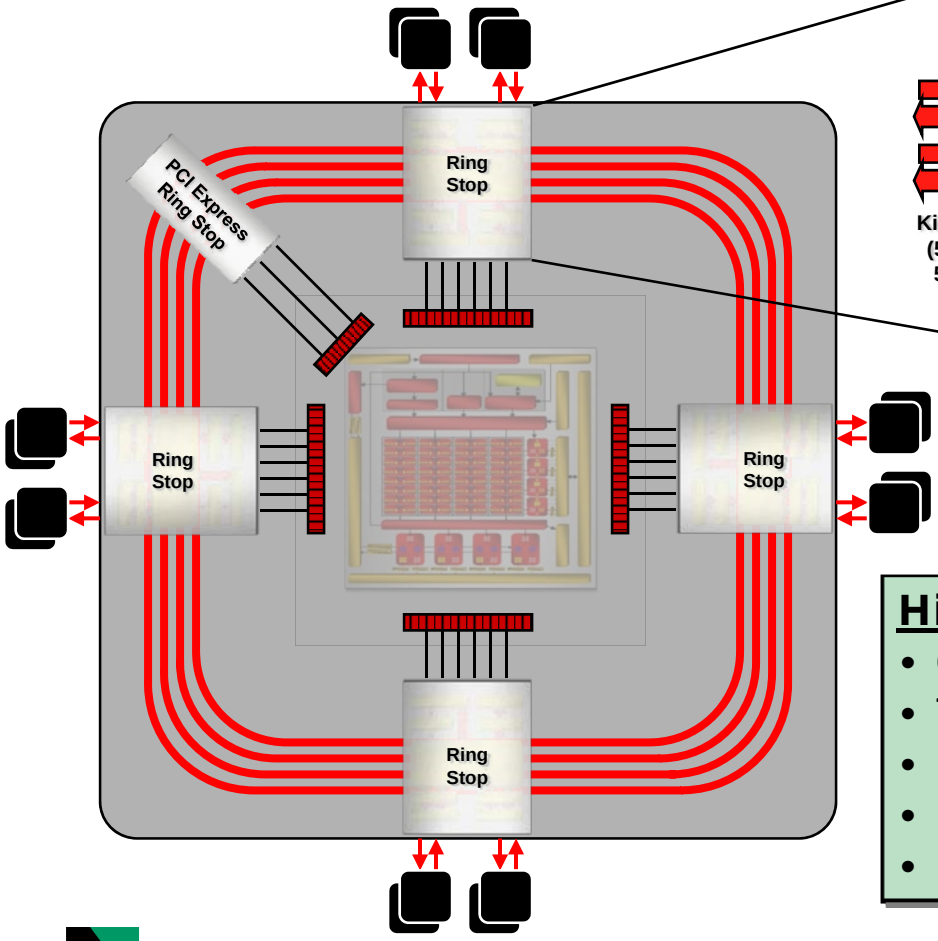
AMD Radeon HD2900 Graphics System

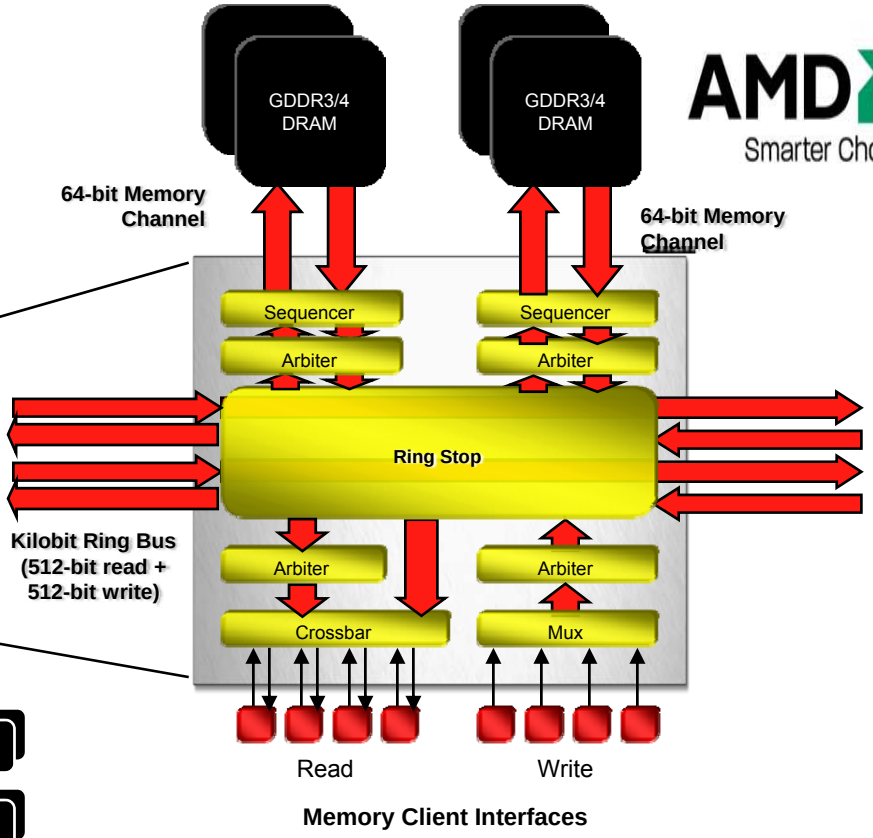


Massive Bandwidth

– World's First 512b Fully Distributed Memory Interface

– New stacked I/O pad design



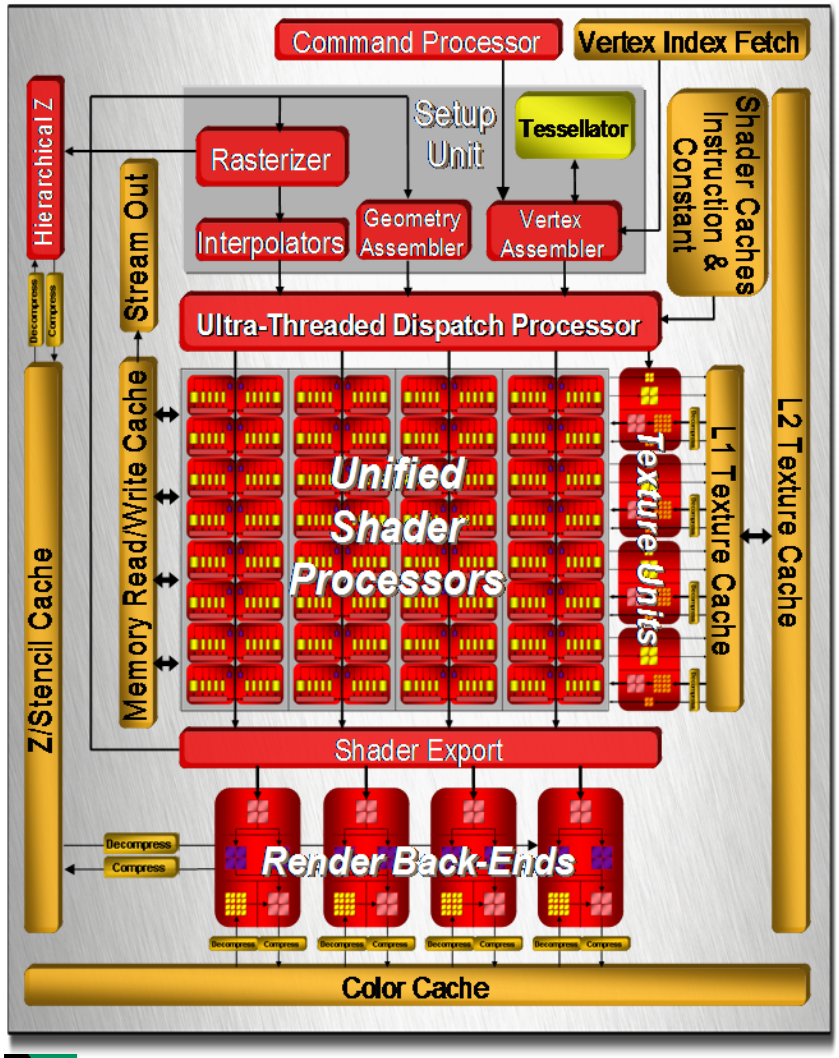


AMD Smarter Choice

Highlights

- Over 100 GB/sec memory bandwidth
- Target achieved via current technology
- Eight 64-bit memory channels
- Kilobit ring bus
- Lower Required Frequencies

AMD Radeon HD2900 Graphics Unit
2nd Generation Unified Shader Architecture



Development from proven and successful XBOX 360 graphics

- New dispatch processor handling thousands of simultaneous threads
- Instruction Cache and Constant Cache for unlimited program size

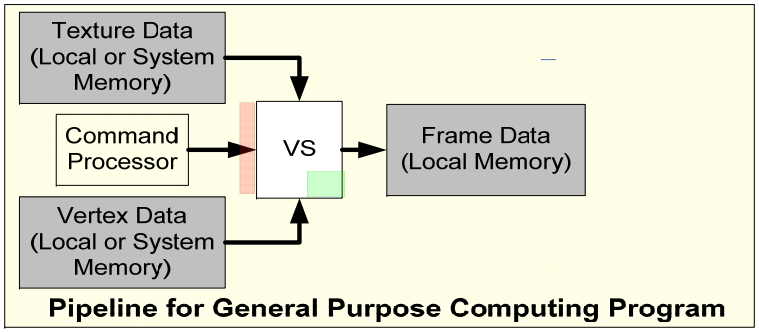
Up to 320 discrete, independent stream processing units

Scalar ALU implementation

- Dedicated branch execution units
- Three dedicated fetch units
  - Texture Cache
  - Vertex Cache
  - Load/Store Cache

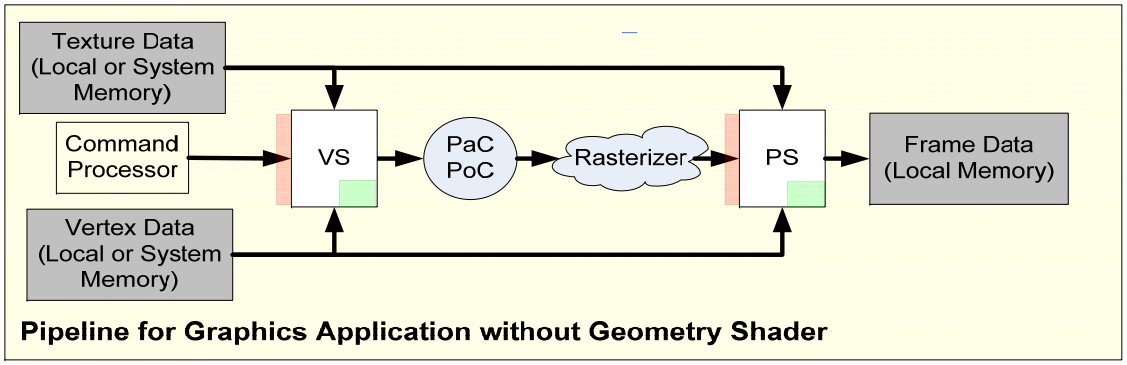
Full support for DirectX 10.0, Shader Model 4.0

Programmer's View of Shader Dataflow



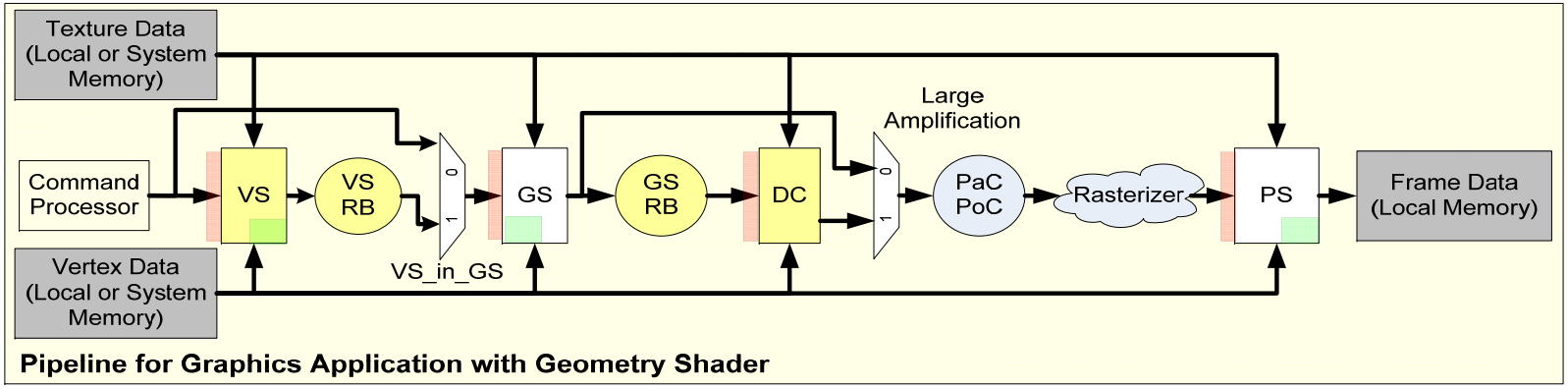
Pipeline for General Purpose Computing Program

VS Vertex Shader
GS Geometry Shader
DC Data Copy Shader
PS Pixel Shader
PoC Position Cache
PaC Parameter Cache
RB Ring Buffer



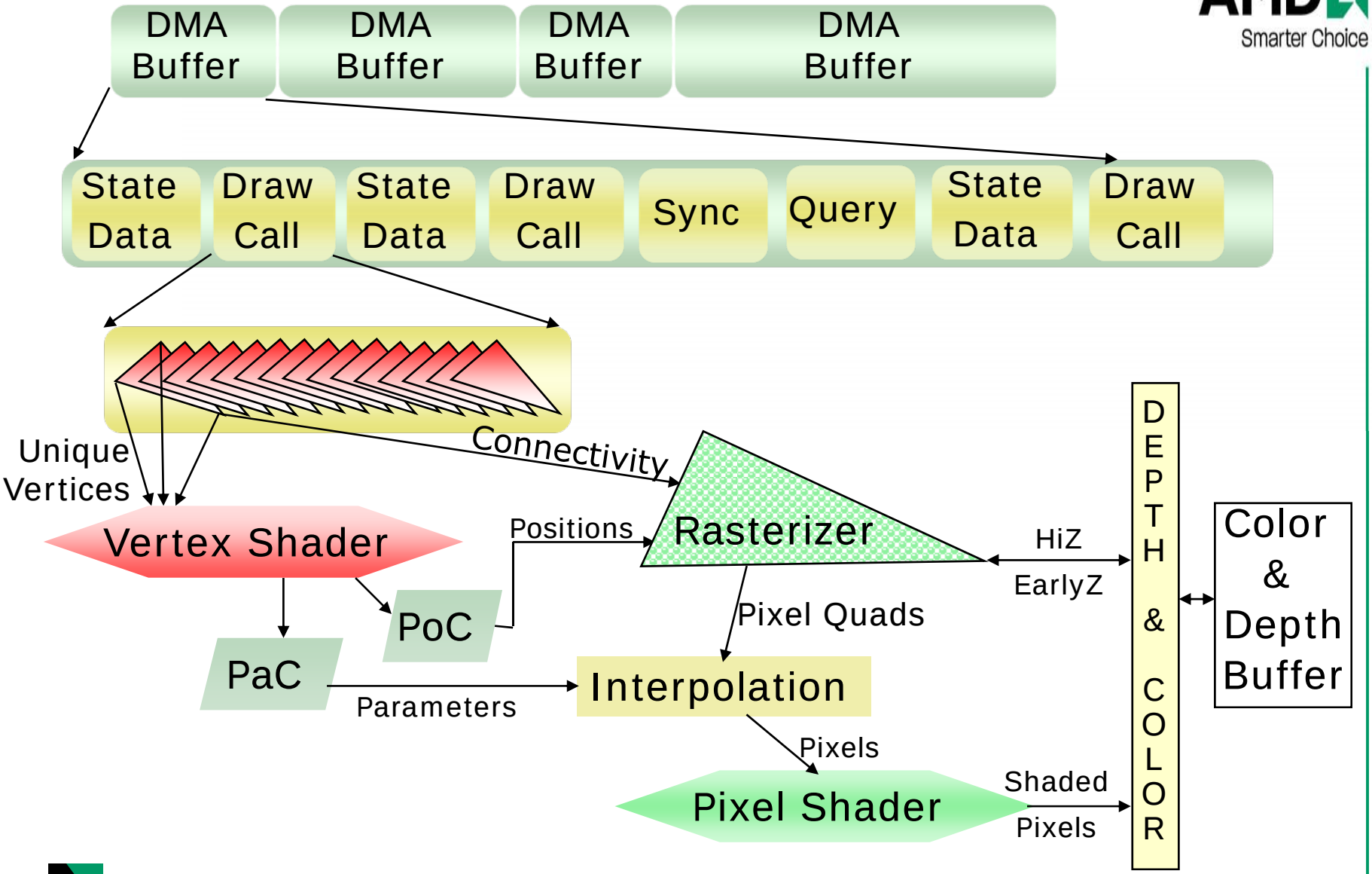
Pipeline for Graphics Application without Geometry Shader

Data R/W Cache
Program Cache
Constant Cache
Setup Stage



Pipeline for Graphics Application with Geometry Shader

Graphics Pipeline Data Flow without GS

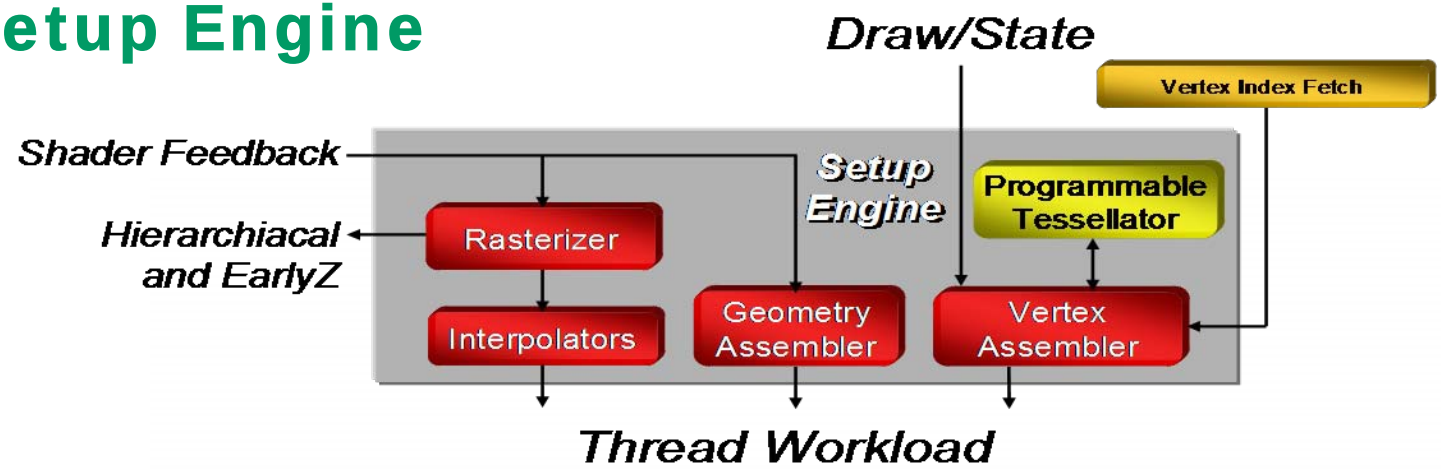


Command Processor

AMD
Smarter Choice

- GPU interface with host
- A custom RISC based Micro-Coded engine
- Memory & Register Read and Write access
- Multiple buffers with dependant fetch latency hiding
- Surface coherency synchronization
- Host interrupt notification system
- Hardware based validation of state data at draw call

Setup Engine



AMD Smarter Choice

Workload preparation for Shader

- Staging to collect data for submission
  - Different arrival/drain rates
  - Different storage requirements
  - Different processing needs
- Submittal Arbitration policies
  - Output Need feedback/Availability/Balance
  - Prevent over-subscription
  - when in doubt favor pixels

- **Vertex workload**
  - – Primitive Assembly & Vertex Reuse
  - – Primitive Tessellation (742Mtri/sec)
  - – Inputs – Index & Instancing Data
- **Geometry Shader Staging**
  - – On\off chip staging
  - – Amplification and parallelism
    - Dependence on SIMD size
- **Pixel Shader Staging**
  - – Rasterization and Interpolation
  - – Vertex/Pixel I/O mappings
  - – Inputs- System variables, z, center, centroid, sample, linear

Design Goals for Unified Shader

AMD
Smarter Choice

- Maximize Performance via ALU utilization
- Provide shared resources for all shader types
- Sustain peak Fetch and I/O Rates
- Provide a common programming language
- Simplify design and verification process
- Enable common tool chain
- Flexibility and Scalability

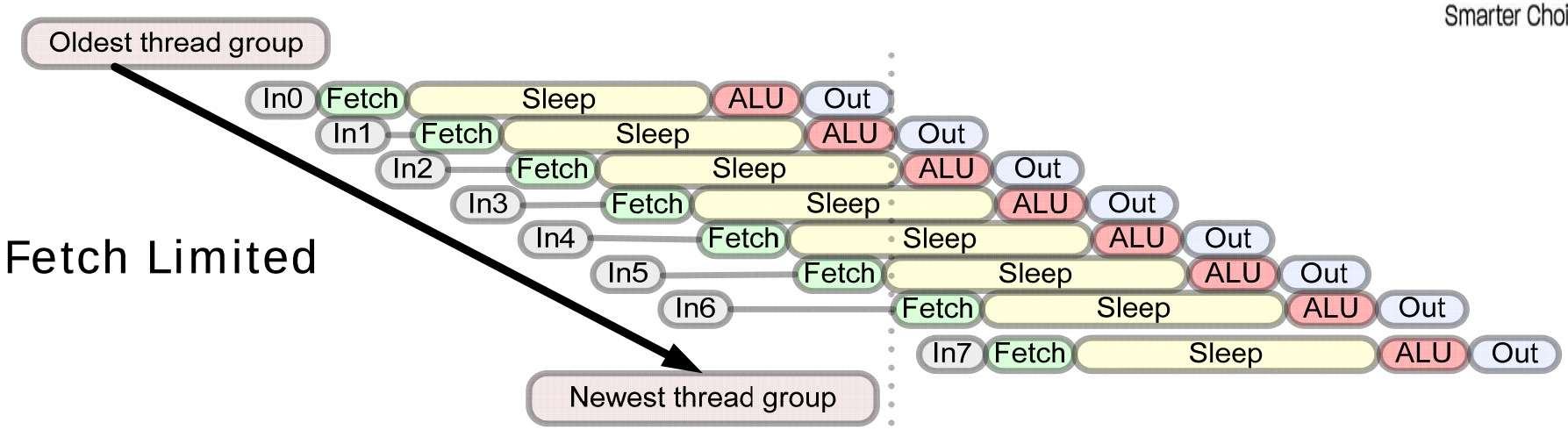
Requirements to meet goals

AMD
Smarter Choice

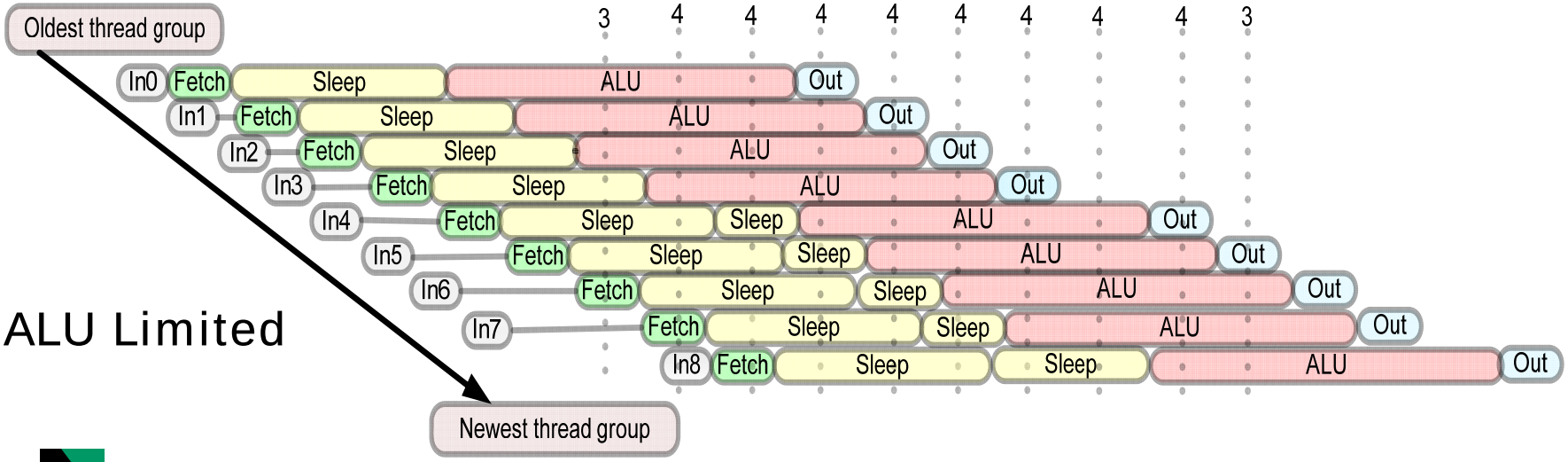
- Hide latency of memory fetches
- Create cache locality to prevent over-fetch
- Prevent resource over subscription
- Arbitrate on age/need to protect bandwidth
- Enable performance scaling for all workloads
- Provide ALU & I/O Ratios for typical workloads
- Interleaved diverse workloads to balance

Simultaneous Multi-Threaded Engine

AMD Smarter Choice



Fetch Limited



ALU Limited

Unified Shader\Stream Processors

AMD
Smarter Choice

Single **I**nstruction **M**ultiple **D**ata

- Each SIMD receives independent ALU instruction stream
- Each SIMD applies instruction stream to multiple data elements

Multiple **I**nstruction **M**ultiple **D**ata

- Multiple SIMD units operating in parallel (Multi-Processor System)
- Distributed or shared memory

Very **L**ong **I**nstruction **W**ord (VLIW) design

- Co-issued up to 6 operations (5 ALU + 1 FC)
- 1.25 Machine Scalar operation per clock for each of 64 data elements
- Independent scalar source and destination addressing

Simultaneous Instruction Issue

- Input, Output, Fetch, ALU, and Control Flow per SIMD

Shader Instructions

AMD
Smarter Choice

VLIW (Very Long Instruction Word), variable length

Control Flow Instructions

- Control branch, loop, stack operations
- Clause launch
- Barriers, Allocation, and Exports

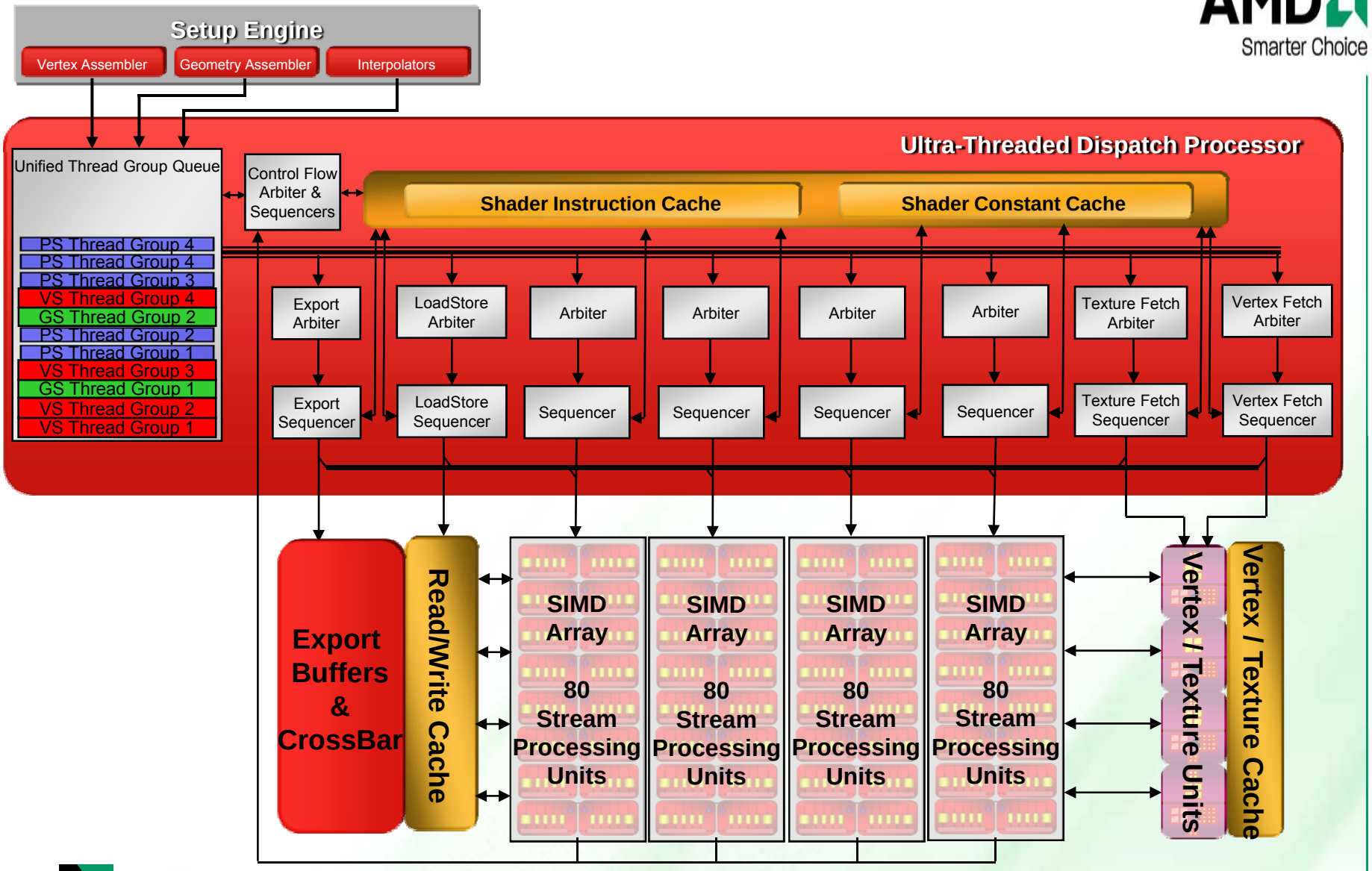
Clause – Set of instructions that executes w/o pre-emption

- ALU Instructions
- Texture & Vertex Fetch Instructions
- Memory Read/Write Instructions

ALU Instruction (1 to 7 64-bit words)

- 5 scalar ops – 64 bits each
- 2 additional words for literal constants

Ultra-Threaded Dispatch Processor



Shader Processing Units (SPU)

AMD Smarter Choice

Arranged as 5-way scalar stream processors

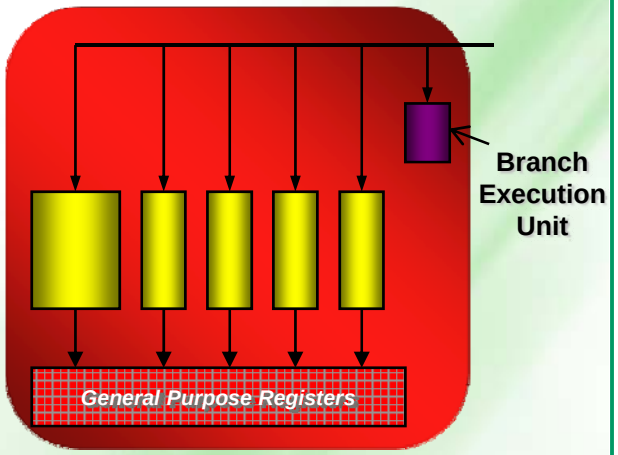
- Co-issue up to 5 scalar FP MAD (Multiply-Add)
- Up to 5 integer operations supported (cmp, logical, add)
- One of the 5 stream processing units additionally handles
  - * transcendental instructions (SIN, COS, LOG, EXP, RCP, RSQ)
  - * integer multiply and shift operations
- 32-bit floating point precision (round to nearest even)

Branch execution units handle flow control and conditional operations

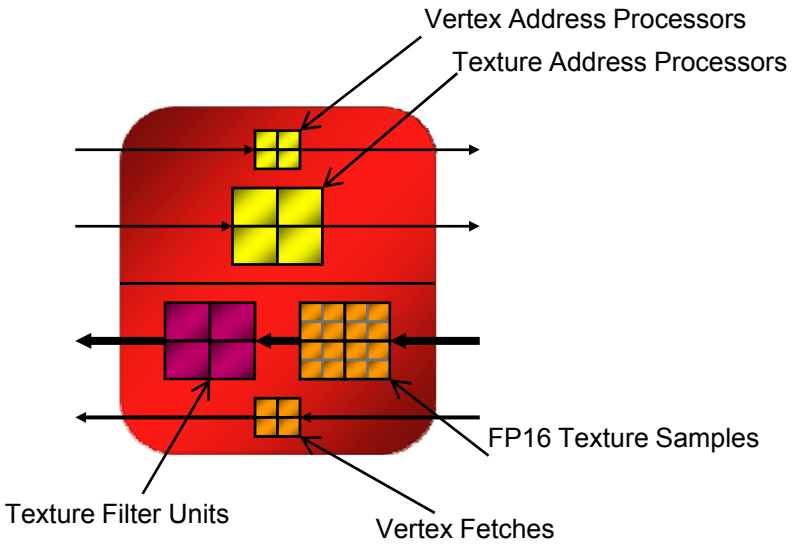
- Condition code generation for full branching
- Predication supported directly in ALU

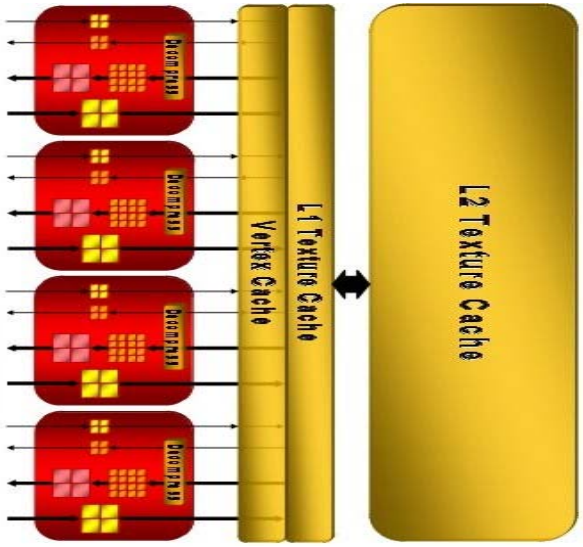
General Purpose Registers

- 1 MByte of GPR space for fast register access



Fetch Unit Design





Fetch units

Fetch Address Processors each

- 4 filtered (fetch neighboring data for filtering)
- 4 un-filtered raw data fetch

20 Samples accessed from cache per clock

4 bilinear filter results per clock (with BW)

- Filter rate for each pixel:

  one 64-bit FP texture result per clock,

  one 128-bit FP result per 2 clocks

Multi-level fetch cache design

L2/L1 cache structures

- Unified 4kb L1 structured cache (unfiltered)
- Unified 32kb L2 structure cache (unfiltered)
- Unified 32k L1 texture cache
- Unified 256KB L2 texture cache

Memory Read/Write Cache

AMD
Smarter Choice

Virtualizes register space

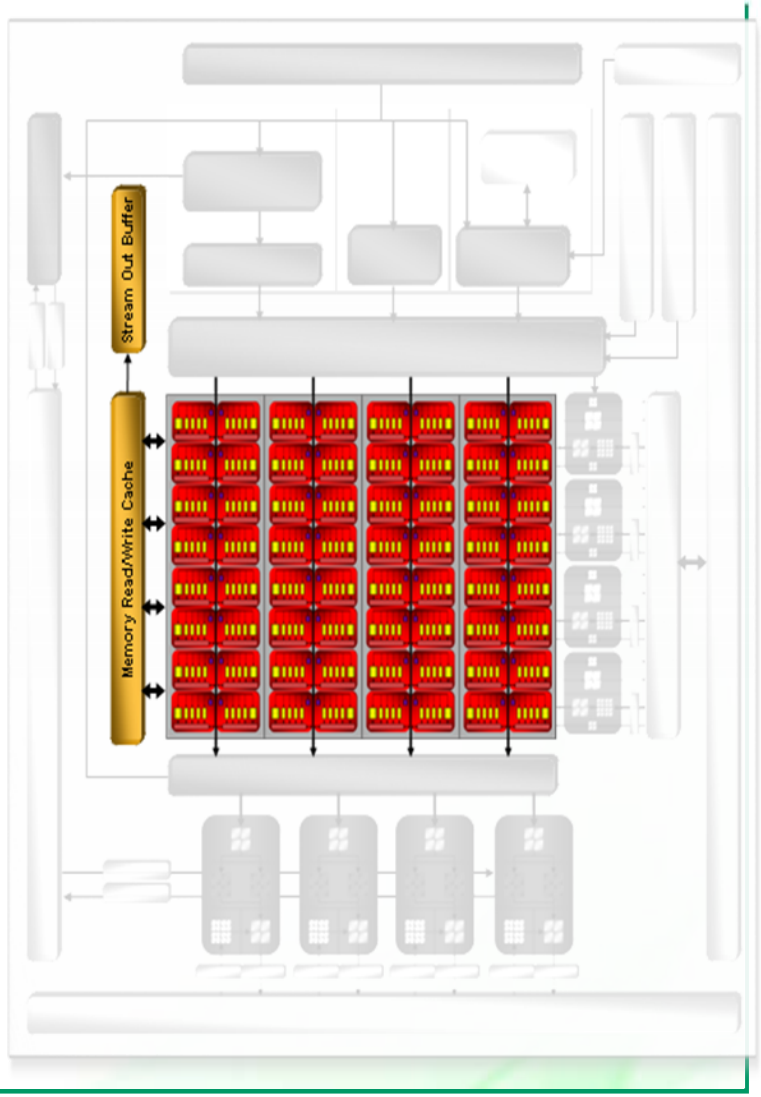
- Allows overflow to graphics memory
- Can be read from or written to by any SIMD (fetch caches are read-only)
- Can export data to stream out buffer
- 8KB Fully associative cache, write combining

Stream Out

- Allows shader output to bypass render back-ends and color buffer
- Outputs sequential stream of data instead of bitmaps

Uses include:

- Inter-thread communication
- Render to vertex buffer
- Overflow storage/output for Geometry Shader data (allowing parallel processing for large amplification)



Render Back-Ends

Alpha testing, Alpha and fog blending

Double rate depth/stencil test

- 32 pixels per clock for ATI Radeon HD 2900

Multi-Sample Anti-Aliasing (MSAA) resolve functionality is programmable

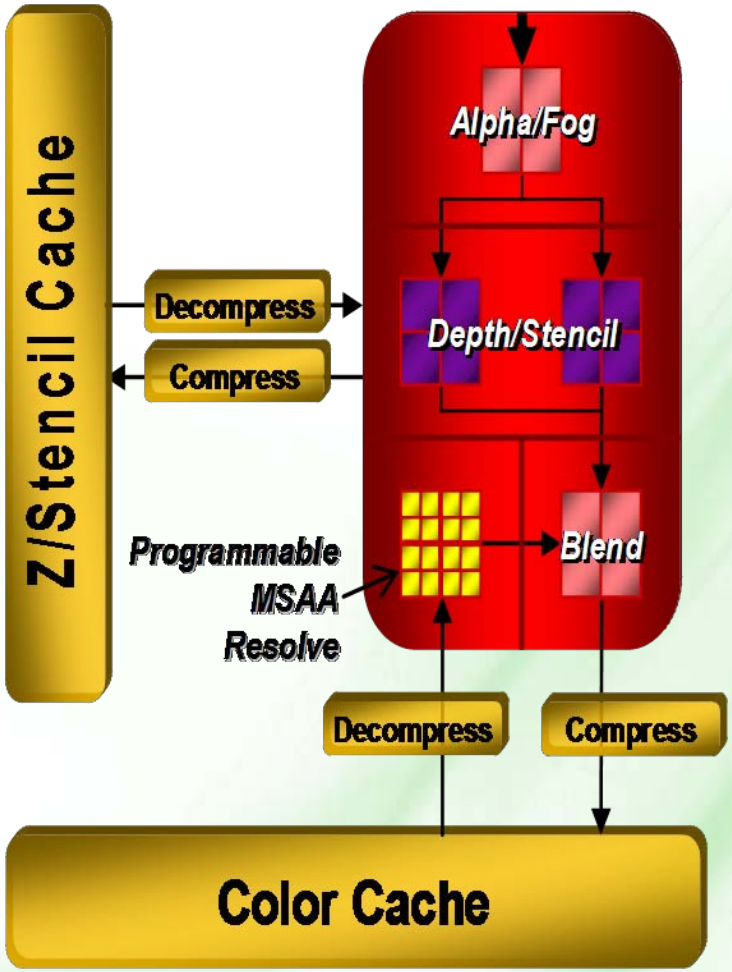
- Makes Custom Anti-Aliasing Filter possible

New blend-able surface formats

- Allows new DirectX10 formats to be displayable
  - 128-bit floating point format
  - 11:11:10 floating point format

MRT (Multiple Render Target) support

- Up to 8 MRTs with MSAA support



A Scalable Family

AMD
Smarter Choice

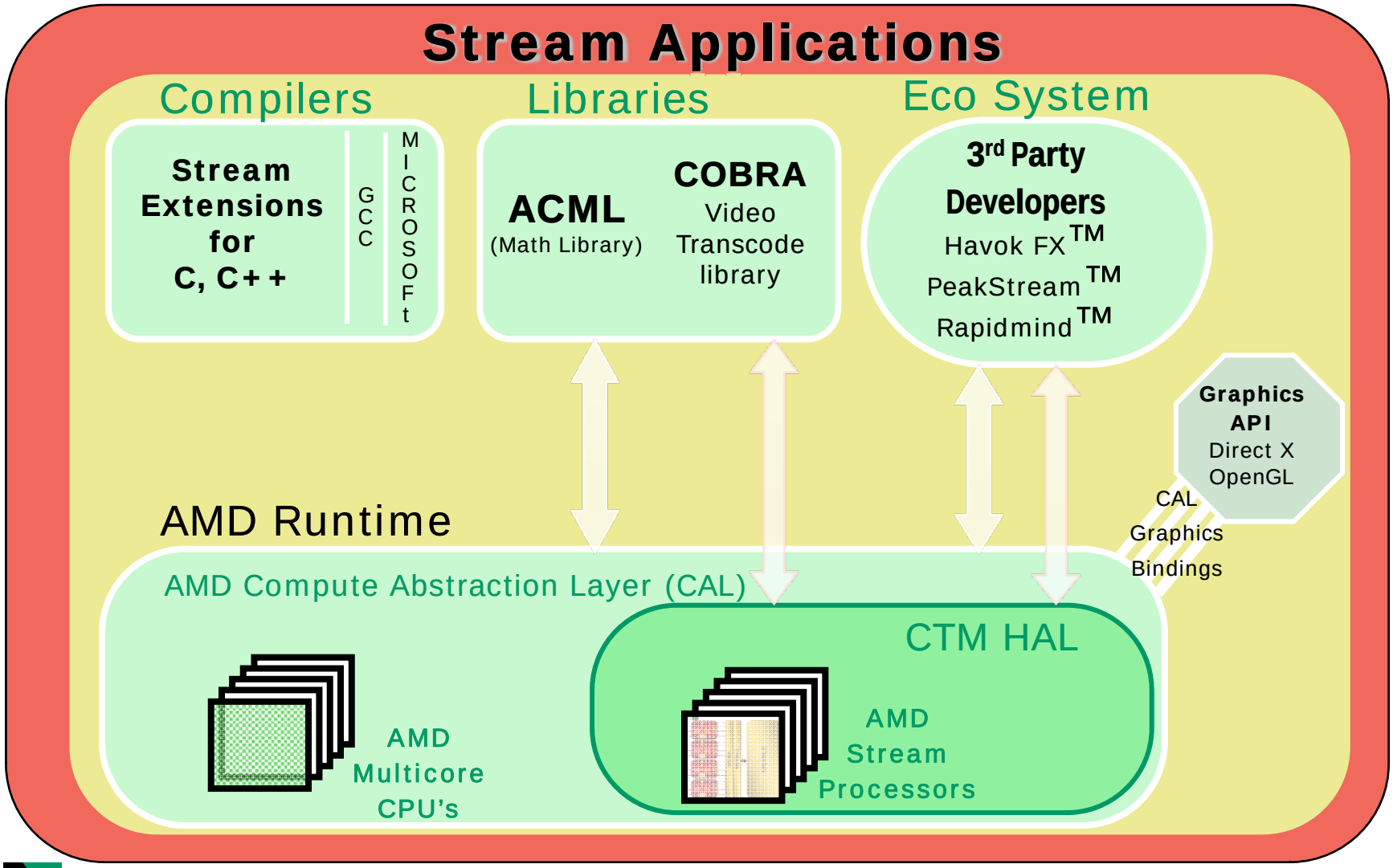
ATI RadeonHD 2900
320 Stream Processors
4 SIMDs
4 Texture Units
4 Render Back-End

ATI Radeon™HD 2600
120 Stream Processing
3 SIMDs
2 Texture Units
1 Render Back-End

ATI Radeon™HD 2400
40 Stream Processing
2 SIMDs
1 Texture Unit
1 Render Back-End
Shared vertex/texture cache

- Designed with a “numbers of” for most elements
- Shader, Texture, Interpolate, Raster Backend Units
- Core functionality exists in all parts
- Target specific cost/performance levels for each part

AMD Accelerated Computing Software



Questions?

AMD
Smarter Choice

Abstract: AMD Radeon HD 2900 Technology

A 2nd Generation Unified Shader Architecture

AMD
Smarter Choice

The internally named R600 Graphics Processor Unit (GPU) developed by Advanced Micro Devices (AMD) will support two distinct types of intensive data parallel compute applications, both high performance 3D graphics processing and general purpose algorithms with large parallel data sets.

The included hybrid Simultaneous Multi-Threaded (SMT) shader core was developed to effectively hide memory fetch latency by interleaving individual vectors of threads grouped in a manner for parallel execution, such that thousands of threads are in flight at any time. Use of both SIMD (Single Instruction Multiple Data) and MIMD (Multiple Instruction Multiple Data) technologies are employed to create the right balance of resources such as Instruction, Constant and fetch capabilities to maintain peak execution rates with high utilization of computation resources when processing large parallel data sets. Vectors of all type of compute (Vertex, Primitive, Pixel, and Compute) are interleaved in this unified shading system so the total compute power is available to each type of processing, while each is used in a cooperative manner to assist in hiding latency. At the lowest level of implementation, instruction level parallelism is utilized along with industry leading compiler technology to schedule a collection of flexible VLIW 5 way scalar processor units to hide pipe latency and achieve maximum utilization rates.

Data and Instruction caches to provide the right balance of inputs for this shading complex has been developed by using a mixture of distributed and unified cache structures. Distributed read only caches for instruction and constants are used to provide a parallel robust set of instruction streams with high re-use of fetched instructions, while providing an interleaved streaming idea for reuse when using very long shader programs. For data fetch read only caches, both a unified texture cache for 1/2/3D texture map fetch and a unified Vertex Cache for array of structure type data fetches is included, each optimized for maximum re-use for that data type and minimizing over fetch.

This GPU is equipped with an enhanced 2nd generation dual ring memory subsystem design with protocol support for DDR3, DDR4, GDDR3, and GDDR4. This dual ring design provides desired bandwidth from any memory channel to desired client.

Also equipped with traditional 3d command buffer fetching/execution capabilities for state, draw and synchronization commands, vertex re-use and rasterization, depth and post shader blend operations.

A thin Hardware/Software Interface layer called CTM (for Close To Metal) will enable general purpose computing access to these compute and fetch capabilities at the lowest level to reach new heights in many areas by enabling data parallel programs to be developed for this device. Applications will be able to harness the extreme available data bandwidths and high arithmetic compute intensity to process parallel data sets in areas such as signal processing, physics, image processing, matrix-matrix multiply, FFT, and convolution

Disclaimer & Attribution

AMD
Smarter Choice

•DISCLAIMER

•The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

•The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

•AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

•AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

•ATTRIBUTION